Classification of Blind Users' Image Exploratory Behaviors Using Spiking Neural Networks

Ting Zhang, Bradley S. Duerstock, *Member, IEEE,* and Juan P. Wachs*, *Senior Member, IEEE*

***Abstract*—Individuals who are blind adopt multiple procedures to tactually explore images. Automatically recognizing and classifying users' exploration behaviors is the first step towards the development of an intelligent system that could assist users to explore images more efficiently. In this paper, a computational framework was developed to classify different procedures used by blind users during image exploration. Translation-, rotation- and scale-invariant features were extracted from the trajectories of users movements. These features were divided as numerical and logical features and were fed into neural networks. More specifically, we trained spiking neural networks (SNNs) to further encode the numerical features as model strings. The proposed framework employed a distance-based classification scheme to determine the final class/label of the exploratory procedures. Dempster-Shafter Theory (DST) was applied to integrate the distances obtained from all the features. Through the experiments of different dynamics of spiking neurons, the proposed framework achieved a good performance with 95.89% classification accuracy. It is extremely effective in encoding and classifying spatio-temporal data, as compared to Dynamic Time Warping and Hidden Markov Model with 61.30% and 28.70% accuracy. The proposed framework serves as the fundamental block for the development of intelligent interfaces, enhancing the image exploration experience for the blind.**

***Index Terms*—Blind or Visually Impaired, Dempster-Shafer Theory, Exploration Procedures, Image Perception, Spatio-temporal Data, Spiking Neural Networks.**

I. Introduction

INDIVIDUALS who are blind often rely on tactual and auditory sensations to perceive images using current sensory-substituted interfaces [1], [2]. However, multimodal sensory substitution for image perception is still a daunting task that requires time, cognitive strenuous, and may require the help of personal assistants [2], [3]. There is a need to develop intelligent systems that can understand users' behavior and provide additional and appropriate assistance to facilitate the perception of images for the blind community. Automatically recognizing users' behavior is a fundamental step towards the development of such systems, which is the focus of this paper.

For people who are blind, the perception of images are constructed based on the visual features gathered from the performed exploration [4]. Blind people use different exploration procedures to better understand specific visual information or features of interest [5], [6]. For example, blind users often follow the contour of an object, to understand its shape and size [6]. Table I summarizes five exploration procedures commonly used by the blind community as reported in the literature and our previous studies, which includes Frame Following (FF), Contour Following (CF), Surface Sweeping (SS), Relative Positioning (RP) and Absolute Positioning (AP) [5], [7]–[9]. In Table I illustrations, blood smear images are used as an example and dotted red lines with arrows represent the trajectory of a user's exploration. Red blood cells are shown in red, and white blood cells are purple.

To automatically recognize the exploration procedures summarized in Table I, a framework using Spiking Neural Networks (SNNs) was developed. These exploration procedures are characterized as distinct spatio-temporal patterns. SNNs have been adopted in various applications to recognize spatio-temporal patterns [10]–[13]. Previous research has indicated exciting results with a relatively high accuracy by giving only a small amount of training samples [14]. This property seems attractive to the studies that benefit the blind community since recruiting high numbers of subjects is challenging [15]. Additionally, SNN-encoded features have shown the capability of early prediction from previous studies, which is well-suited for intelligent systems [11].

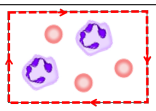
Previous studies have focused on training a single SNN to characterize the spatio-temporal patterns with only one feature, such as the traversed pixels or the angles of movement of the trajectories [11], [14]. However, exploration procedures have to be characterized by multiple features that are invariant to scale, rotation and translation. For example, the trajectories of procedure RP and AP show similar features in terms of directions of movements. To distinguish them, it is necessary to include the morphological differences of objects on the image that are related to the procedures. Procedure RP interacts between two objects, while procedure AP interacts between one object and the boundary of the image. Therefore, a multimodal SNN approach was developed in this paper to encode the spatial and temporal characteristics of each pattern. Multiple SNNs were adapted and trained to encode these features. With this unsupervised encoding of features using SNNs, distance-based classification and decision fusion techniques are applied to distinguish the exploration procedures.

The contribution of this paper is three-fold: (1) developed a computational framework that recognizes five exploration procedures frequently used during image exploration for the blind; (2) conducted experiments to measure the performance of the proposed approach; (3) performed analysis to evaluate the effect on classification performance of different dynamics of SNNs.

T. Zhang is with the school of Industrial Engineering, Purdue University, West Lafayette, IN, USA. B. S. Duerstock is with the School of Industrial Engineering and Weldon School of Biomedical Engineering, Purdue University, West Lafayette, IN, USA. * J. P. Wachs is with the school of Industrial Engineering, Purdue University, West Lafayette, IN, USA (correspondence e-mail: jpwachs@purdue.edu).

Manuscript received April 2, 2019.

TABLE I
SUMMARY OF EXPLORATION PROCEDURES.

Exploration Procedure	Description	Illustration
Frame Following (FF)	Trace the boundary of the image to obtain the image size.	
Contour Following (CF)	Trace the boundary of objects on the image to learn the size and shape of objects.	
Surface Sweeping (SS)	Back-and-forth movement inside objects to learn the features of objects.	
Relative Positioning (RP)	Back-and-forth movements between objects to obtain their relative locations.	
Absolute Positioning (AP)	Back-and-forth movements between objects and the image boundary to obtain their absolute locations on the image.	

The rest of the paper is organized as the following. Section II summaries the state-of-art. Section III explains the proposed framework for procedure recognition. The evaluation of the proposed methodology and results of the experiments are illustrated in section IV. Section V discusses the potentials for this work. Section VI gives the conclusion and presents future work.

II. Related Work

In this section, we reviewed current assistive technologies for people who are blind to perceive images, and then discussed various exploration procedures.

A. Assistive Technologies for Image Exploration

There has been a substantial interest in developing assistive technologies to allow visual information be accessible to blind individuals ranging from low-tech tactile papers to high-tech real-time sensory substitution systems [16]–[19]. The state of the art in image exploration techniques can be summarized into two catogories: exploring images physically or digitally.

Among physical image representations, tactile graphics is the most common approach to deliver two-dimensional visual information to individuals who are blind or visually impaired (BVI), such as diagrams, pictures and charts. Despite of its widespread use, tactile graphics cannot be used to deliver complex visual information and often requires additional description from human assistants [20]. In addition to tactile paper, 3D printing technology became an alternative as 3D printers becomes more affordable [21]. In spite of its apparent superiority to represent more complicated visual information than tactile paper, studies indicated that 3D printed tactile images require fine adjustment of parameters and 3D modeling techniques to convert 2D images into 3D objects [22].

Complementary to the physical representation of images, force feedback devices have drawn much attention for blind individuals as another modality that supports digital image exploration [2], [23], [24]. Instead of using fingers, blind individuals can feel the virtual graphic rendered by the computer via the motion of a haptic device across the image. Tactile feedback is provided according to the position of the haptic device. Yu and Brewster developed a multimodal system that people can use to explore bar graphs using a haptic device, together with auditory feedback [25]. Experimental results indicated significant improvement in the understanding of bar charts through the haptic interface compared to traditional tactile diagram. In addition to this, interactive systems have been developed using haptic devices and sound to facilitate the learning of visual concept, such as astronomy [26] and maps [27]. Students reported better understanding of the concepts and increased interest in the learning of complex visual concepts.

B. Exploration Procedures

As opposed as to using vision to perceive images, exploring visual information through touch is more difficult and complicated. It imposes cognitive load as well as effective (well-developed) procedures [28]. Studies focusing on what type of procedures are adopted by the blind community have been conducted in two major types of visual properties: local and global visual information.

Lederman & Klatzky [4], [6], [29] summarized six exploration procedures when internal properties of objects are acquired including dimensions, texture, hardness, weight, and shape. For example, “lateral motion” or “surface sweeping” referred in [30] is a back-and-forth movement across a small area within the object’s surface. People use this procedure to understand the texture, interior structure of an object. “Contour following” is another often-used procedure to understand the exact shape of an object. Different from “surface sweeping”, it is a motion that focuses on the boundaries of objects. Vinter et al. summarized seven procedures from studies with children who are BVI using elevated 2D tactile images [5]. The “contour following” procedure was the most common found in their study, and was highly correlated with the recognition of object’s shape and size. “Surface sweeping” was the second most common procedure found in their study.

Besides understanding the internal properties of objects, measuring the spatial relations between objects is another challenging task for BVI individuals. In the experiments conducted [31] with children with visual impairments, participants were asked to explore a circular board with a number of objects placed on it, and then replicate the setup on a different board. Results revealed that participants finished the task with higher accuracy when they applied “relative positioning” and “absolute positioning” procedures. Other works conducted with adults who are BVI confirmed such results [32].

III. Methodology

Trajectory of users’ exploration behavior was collected during trials involving subjects exploring images using a previously developed multimodal interface [2]. In our proposed framework, a sequence of multidimensional feature vectors

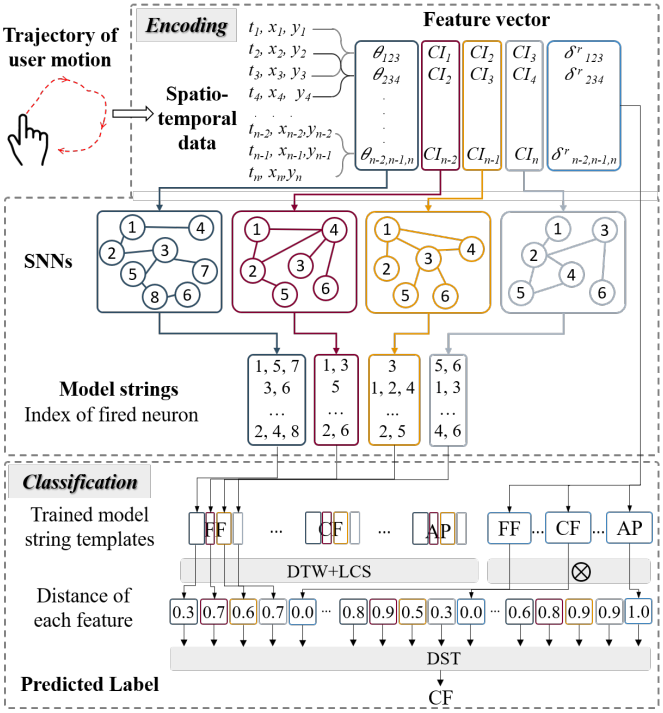


Fig. 1. Proposed framework to learn exploration procedures. The trajectory of user motion is characterized as spatio-temporal data and encoded as model strings through the training of SNNs. Classification of the exploration procedures are then performed by computing distances between samples.

was extracted from each sample of an exploration procedure. The features represent the directions of movements, context of the trajectory and users' frame of reference during exploration. The features representing the angles and context information of trajectories have numerical values, while the feature representing users' frame of reference is a logic variable (e.g. has the reference been changed or not). The numerical features are further encoded through trained SNNs. These features are fed to the SNNs for training acting as the stimulus of the network. Once the SNNs are trained, the characteristic responses of each sample to the SNNs are encoded as model strings, serving as the templates for each sample. Those templates are further used for classification, together with the logic feature (indicating the switch of reference points during the interaction). For classification, we adapted a modified version of Dynamic Time Warpping (DTW) with Longest Common Subsequence (LCS) as the similarity measurement between model strings. Dempster-Shafer Theory (DST) was then applied at last to merge the beliefs from multiple features into a final decision (i.e. the predicted type of the sample). The proposed framework is illustrated in Fig. 1.

A. Data Collection

Trajectory data were collected from 10 blind-folded human subjects exploring 12 blood smear images using a haptic-based multimodal image exploration interface. In previous studies, we experimented with both blind and blind-folded subjects with no significant difference in performance [2]. The multimodal interface consists of a haptic controler (Force Dimentions Omega 6), a vibration tactor (Engineering Acoustics, Inc) and a computer speaker for audio feedback. The sample rate of the multimodal interface is 10 Hz with the spatial resolution of less than 0.01 mm and workspace of ø 160 mm by 110 mm. Recruited participants included 5 females and 5 males with age ranging from 18 to 30. Participants were asked to explore a blood smear image using the multimodal interface and replicate the image with 3d printed cell models. The replicated image indicated their understanding of the image.

Exploration procedures were used as samples and were manually annotated. The procedures were segmented from a longer data trajectory corresponding to a full image routine. In total, the dataset included 168 samples for FF, 803 for CF, 207 for SS, 238 for RP and 483 for AP. Twenty samples of each procedures were separated from the dataset working as the training set, while the rest of the samples were used for testing. Fig. 2 shows an example of Absolute Positioning (AP) in three iterations. The different shades of red indicate the elapsed time of trajectory. The lighter shades happened earlier than darker shades.

B. Temporal Representation of Data

Each sample of trajectory data is a sequence of time and cursor's 2D position on the image, represented as a tuple (t_i, x_i, y_i), where i ranges from 1 to the length of the sample. The procedures are then extracted from the trajectory data as spatio-temporal patterns, that have different spatial appearance over time. Three types of features were computed to characterize the trajectories in a fashion that is translation, rotation and scale-invariant: the "angle of motion", the "context of cursor's position" and the "switch of reference point".

1) Angle: For any two consecutive time frames t_{i-1} and t_i, the vector v_i representing the direction of movement is obtained following (1). The angle of motion $\theta_{i-1,i,i+1}$ is then computed as the angle between two consecutive vectors v_i and v_{i+1} following (2).

$$v_i = (x_{i-1} - x_i, y_{i-1} - y_i) \tag{1}$$

$$\theta_{i-1,i,i+1} = \cos^{-1} \frac{v_i \times v_{i+1}}{\|v_i\| \|v_{i+1}\|} \tag{2}$$

2) Contextual Information: To model user behavior, it is crucial to include the context of the interaction as it partially determines the procedure that a user will perform. For instance, contour following (CF) is defined as a procedure

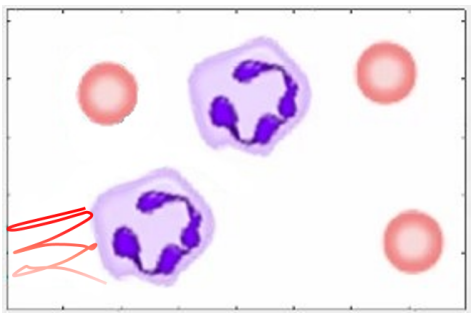


Fig. 2. An example of Absolute Positioning (AP) performed by participants.

TABLE II
CONTEXTUAL INFORMATION ASSIGNMENT.

Index	Contextual Information
1	Background
2	Object contour
3	Object Inside Area
$4 \sim 8$	Pixels from 1 to 5 away outwards from the object contour
9	Image boundary
$10 \sim 13$	Pixels from 1 to 4 away from the boundaries of an image

that user follows the boundary of an object. Despite of the variant shapes of objects, the traversed pixels in CF should belong to the contour of an object that is recognized through image processing techniques. Therefore, shape- and location-invariant features are defined by creating an index map of the image to be explored. An index is assigned for each unique contextual information (CI_k) for a given time frame k, following the rules explained in Table II.

3) Reference Switch: When people explore a visual landscape, such as an image, using only tactile information, features and objects are used as reference points. This can be accomplished by applying image processing techniques to segment the different objects, whereas each object gets unique IDs (ranging from object 1 to n, where n is the number of objects in the image) [2]. Once the objects are recognized, a reference point is defined as the object that the user is in contact with, including the image boundary and object number i ($i = 1 \dots n$). The last reference point r_i is annotated for each time frame t_i. The switch of reference point (δ^r) is computed based on three consecutive time frames, following (3).

$$\delta^r_{i-1,i,i+1} = \begin{cases} 0, & r_{i-1} = r_i = r_{i+1} \\ 1, & \text{otherwise} \end{cases} \tag{3}$$

Next, a feature vector (4) for the spatio-temporal data is extracted using three consecutive time frames t_{i-1}, t_i and t_{i+1}, where, the angle $\theta_{i-1,i,i+1}$ is computed using (2), the context index (CI_i) is calculated for each time frame t_i following Table II and the reference switch $\delta^r_{i-1,i,i+1}$ is obtained using (3).

$$f = (\theta_{i-1,i,i+1}, CI_{i-1}, CI_i, CI_{i+1}, \delta^r_{i-1,i,i+1}) \tag{4}$$

C. Spiking Neural Network

SNNs are trained to encode the numerical features in (4), including the "angle of motion" and the "context information". As defined in (3), it may be cumbersome to encode $\delta^r_{i-1,i,i+1}$ using SNN as it is a logic variable with Boolean values, so this feature will be used for classification directly. Therefore, there are four SNNs trained in this work, each corresponding to feature $\theta_{i-1,i,i+1}$, CI_{i-1}, CI_i and CI_{i+1}. The four SNNs have different number of neurons due to the range of features, but same configuration for other parameters.

1) Network Configuration: In an SNN, neurons are mutually connected through the synapses, and can be configured with the number of neurons, the connectivity among neurons, the conduction delay of each synapse, and the weights of synapses, as defined in Izhikevich [33]. Izhikevich's model was used in this work considering its ability to simulate different neuron dynamics while requiring relatively small computational power, compared to other widely used models, such as "Integrate-and-Fire" and "HodgkinHuxley" [34]. There are two types of neurons in an SNN, the excitatory and inhibitory neurons, of which the amount has a ratio of 4:1. The number of neurons for each of the four SNNs is then determined based on the range of the feature it is encoding and discussed in the next section. These neurons are mutually connected and the four SNNs developed in this work have a connectivity of 10% among all neurons. The synapses can have different conduction delay that partially determines the firing pattern of neurons. The differences between firing patterns of input data are then used for classification. The conduction delay is randomized from 1 ms to 20 ms for each synapse. The synaptic weights are configured as $+6$ for excitatory neurons, and -5 for inhibitory neurons initially. The weights are then updated using the rule of the spiking dependent plasticity (STDP) [33]. Based on the time-locked firing patterns of neurons, the STDP rule boosts or degrades inter-neuron connections by increasing or decreasing the synaptic weights. Fig. 9 in Appendix illustrates an example SNN.

2) Number of Neurons: The number of neurons for an SNN is dependent on the range of the feature. The degree of feature angle $\theta_{i-1,i,i+1}$ ranging from 0 to 180 inclusive, is divided into 19 intervals. Five excitatory neurons are allocated for each interval, resulting in 95 excitatory neurons. Twenty-three inhibitory neurons are allocated as the ratio between excitatory and inhibitory neurons is 4 to 1 as discussed above. The SNN encoding feature angle has a total of 118 neurons, with 95 excitatory and 23 inhibitory neurons.

For feature contextual information (CI_i), the index ranges from 1 to 13 as defined in Table II. Similarly to the configuration for feature angle, each index is allocated with 5 excitatory neurons, thus leading to a network of 81 neurons with 65 excitatory and 16 inhibitory neurons.

3) Network Training: The input data (feature Angle and CI) are fed into the network following the sequence of feature vectors defined in (4) by stimulating the corresponding neurons with a 20 mA current. Since there are 5 neurons allocated for each interval of feature values, the neurons are stimulated one by one with an interval of 1 ms. For instance, if the value of feature CI at a certain time frame is associated with neurons 1 to 5, then neuron 1 is firstly stimulated with a 20 mA current. After 1 ms, neuron 2 is then stimulated, followed by neuron 3, 4 and 5 at a 1 ms interval. A training sample with n time frames will have a pattern with length of $5n$ ms fed into the SNN. Fig. 3 shows an example of how neurons were fired given the input sequence of feature angle. The raster plot on the top indicates the firing patterns of neurons, while the bottom plot shows the input values of feature angle over time.

To learn the 5 types of procedures, m sets of training samples were used where each set of training samples contains one sample from each class . Therefore, there are 5 samples in one training set. The total number of training samples is $5m$.

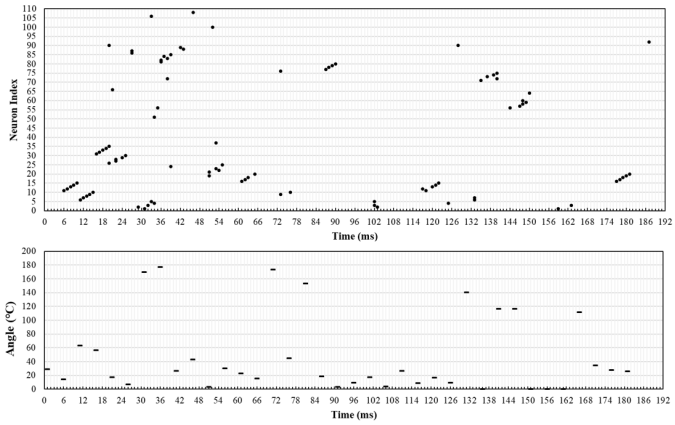


Fig. 3. Firing pattern of neurons. The top raster plot shows the fired neurons at each time stamp and the bottom plot shows the value of input data at each time stamp.

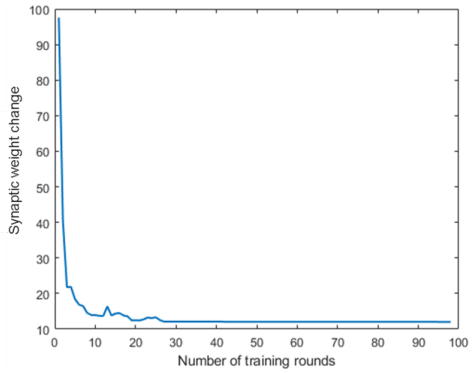


Fig. 4. Examples of synapse weights fluctuation. The synaptic weights remain stable after 30 rounds of training.

One round of training involves feeding these sets of training samples into the network one by one. The STDP rule is used during the training process to update the synaptic weights for all neurons. Till the end of each round of training, the summation of the synaptic weights of all neurons is calculated. For two consecutive rounds of training, the differences between the synaptic weights, ΔW, are computed following (5), where w_p^i is the weight of synapse p at training round i, and the SNN has a total number of n synapses. The training is completed when synaptic weights remain stable. Fig. 4 shows an example of the fluctuation of synaptic weights over different rounds of training. The training of the network is completed around 30 rounds.

$$\Delta W = \sum_{p=1}^{n} \|w_p^i - w_p^{i+1}\|^2 \tag{5}$$

D. Classification

The features encoded by SNNs, including the angle and CI are used together with feature "reference switch" δ^r to classify the various procedures. A distance-based metric is used to measure the similarities between samples. For SNN-encoded features, a modified dynamic time warping (DTW) is used to compute the distance, while for feature δ^r, the distance is computed as the difference between δ^r over time. Then, Dempster-Shafer Theory (DST) is applied to determine the type of procedure being used by combining the distances obtained from all the features.

1) Representation of SNN-encoded Features: Features encoded by SNNs are represented as *model strings* that are sequences of *characters*. The length of model strings varies according to the length of the input sample. One *character* contains the indices of fired neurons (voltage $\geq 30\ mV$) at a time stamp of the training process [11]. For example, {(2,6,7), (3), (5,6)} is a *model string* that contains three *characters*. (2,6,7) is a *character* that indicates neuron 2, 6 and 7 are fired at the same time. The *characters* in a *model string* are formed one by one following the order of time. Considering the same example showed above, neuron 5 and 6 fired at the same time after neuron 3 fired. *Model strings* are built by stimulating the trained SNN without updating the synaptic weights of neurons.

2) Distance Function: As samples of data have different number of time frames, DTW is used to align the sequences and compute the distance between the *model strings*. The length of Longest Common Subsequences (LCS) is used as the distance function of DTW to compute the distance between *characters*. Since LCS computes the similarities of two *characters* instead of distance, its negative value is used for DTW.

To calculate the distance between two samples a and b in the logic feature δ^r dimension, the difference of summation over time is used, follow (6), where sample a has p time frames and sample b has q time frames.

$$d_{a,b}^r = \begin{cases} 0.0, & \sum_{t=1}^{p} \delta_{a,t}^r = 0 \text{ and } \sum_{t=1}^{q} \delta_{b,t}^r = 0 \\ 0.0, & \sum_{t=1}^{p} \delta_{a,t}^r > 0 \text{ and } \sum_{t=1}^{q} \delta_{b,t}^r > 0 \\ 1.0, & \text{otherwise} \end{cases} \tag{6}$$

In (6), condition 1 indicates there is no reference switch in both sample a and b, while condition 2 means that user's reference was switched in both samples a and b. Therefore, there is no difference between the two samples in terms of "reference switch".

3) Decision Fusion using Dempster-Shafer Theory (DST): With a 5-dimension feature vector defined in (4), DST is used to determine the label of the observation by combining the knowledge obtained from each individual feature. Compared to feature fusion, this kind of decision fusion scheme is more flexible and can easily include new features or exclude unnecessary ones [35], [36]. In DST, to classify a sample as label y, the contribution of each feature k is characterized by the Basic Belief Assignment (BBA) function $m_k(y)$, where $k = 1, 2..., 5$ and $y \in \{FF, CF, SS, RP, AP\}$ in this study. For feature k, the value of the BBA function $m_k(y)$ is determined using the average distance between the testing sample and procedure y's trained templates. The Dempster's Rule of Combination (DRC) is then applied repetitively to calculate the joint BBA function $m_{1,2}(\cdot)$ from two individual BBA functions $m_1(\cdot)$ and $m_2(\cdot)$.The joint BBA function obtained from each iteration is then combined with one of the remaining BBA functions of a single feature, until all features are combined [12].

The complete classification procedure is summarized in Algorithm 1 below.

Algorithm 1: Classification with modified DTW and DST

Input: Testing model string t of length $n_t(c_1, c_2, \ldots, c_n)$
Output: **label of** t: predicted type of exploration procedure

for each training sample x of length n_x **do**
  for each feature channel k **do**
    for each character c_i in t **do**
      for character c'_j in x between the window **do**
        $similarity = -$ LCS (c_i, c'_j)
        $cost(i,j) = similarity +$
        $\min(cost\{(i,j),(i+1,j),(i,j+1)\})$
      end for
    end for
    $distance(t,x,k) = cost(nt+1, nx+1)$
  end for
end for
for each feature channel k **do**
  for all training templates x of label y in S **do**
    $m_k(y) =$ mean $\sum_{x \in S} distance(t,x,k)$
  end for
end for
return label of $t \leftarrow$ DRC$(m_1, \ldots, m_k)$

IV. EXPERIMENTAL RESULTS

Experiments were conducted to evaluate the proposed framework using real world data collected from human participants.This study (protocol number: 1207012484) is approved by the institutional review board at Purdue University and informed consent was received from all participants.

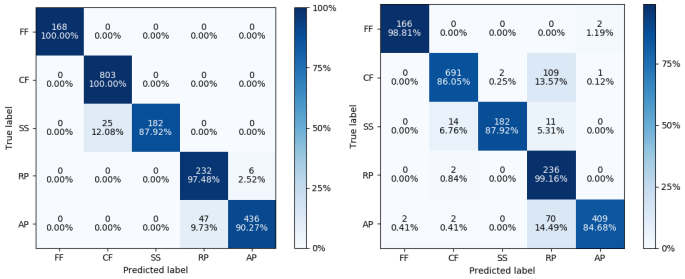
A. Evaluation

A 10-fold cross-validation was performed to evaluate the proposed framework, where in each fold, a leave-n-subject-out practice was utilized. The value of n depends on the number of training samples. For example, when one sample is used for training, the user of the training sample is eliminated from the testing set which results in $n = 9$. Otherwise, when there are 9 training samples, 9 participants' samples are used for training, so the remaining one subject's data are used for testing ($n = 1$). The proposed framework has an average classification accuracy of 95.89% with 18 training samples for each exploration procedure type. Fig. 5a shows the confusion matrix, where rows represent ground truth and columns represent predicted labels. The precision rate, recall rate and F1 score were also computed for each class of exploration procedure according to the one-vs-rest basis (Table III).

More analysis was conducted to understand the effect on classification accuracy when the SNNs were trained with different number of samples. Fig. 6 shows the average classification accuracy with variance over different number of training samples. The accuracy reaches 94% with 5 training samples.

B. Time Complexity Analysis

The proposed classification algorithm is an instance-based approach where the testing sample is compared with all the trained templates. It has a time complexity of $O(mN)$ where m is the number of classes and N is the number of trained templates for each class. With the increasing of number of training samples, classification takes longer to complete.



(a) (b)

Fig. 5. Confusion matrix of classification accuracy for (a) the proposed framework with 95.89% accuracy; (b) SNN with NHNF descriptor with 88.68% accuracy.

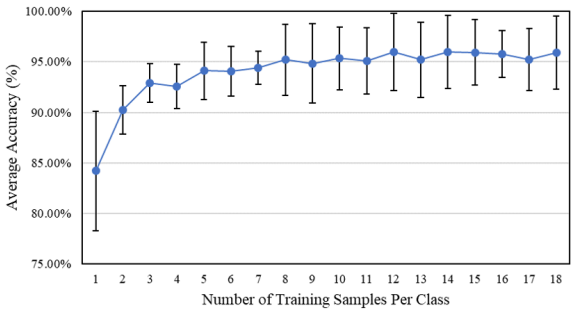


Fig. 6. Classification accuracy over different number of training samples.

Instead of using distance-based metrics, a classification scheme that requires constant time complexity $O(1)$ was developed in [37] for SNN encoded features. A Normalized Histogram of Neuron Firings (NHNF) is firstly computed and then trained with a support vector machine (SVM) for classification. The NHNF approach was tested in this paper by concatenating the normalized histogram of the logic feature "reference switch" with the NHNFs of the four SNN-encoded features. Classification was then performed by training this concatenated histogram using a SVM. The 10-fold leave-one-subject-out cross-validation was performed for this comparison. Compared with the proposed approach with 95.89% accuracy, faster classification was achieved by sacrificing the accuracy to 88.68% with the NHNF approach. Fig. 5b shows its confusion matrix.

C. Comparisons with DTW and HMM

The proposed framework was validated by comparing it with other popular algorithms for time-series data classification, such as DTW [38] and HMM [39]. Without the encoding of features using SNNs, DTW and HMM were trained and

TABLE III
THE PRECISION RATE, RECALL RATE AND F1 SCORE OF THE RESULTS.

	FF	CF	SS	RP	AP
Precision	1.00	0.97	1.00	0.83	0.99
Recall	1.00	1.00	0.88	0.97	0.90
F1 score	1.00	0.98	0.94	0.90	0.94

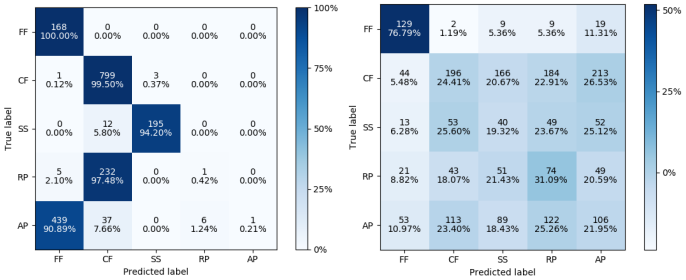
tested using the raw features extracted from data in (4). In these comparisons, 18 training samples were used for each type of procedure. The leave-one-subject-out cross-validation was also performed.

Comparing with DTW, Euclidean distance was used to calculated the differences between samples in terms of feature angle since it is a continuous value, while the LCS used in sectionIII-D2 was used as the distance function for feature context index and reference switch. Each testing sample was compared with 18 training samples for each type of procedure. The DTW confusion matrix shows an accuracy of 61.30% (Fig. 7a). It was observed that Frame Following (FF), Contour Following (CF) and Surface Sweeping (SS) were mostly recognized, while DTW failed with the other two types of procedures. Relative Positioning (RP) were likely recognized as CF because both procedures were related to the objects on an image. In contrast, Absolute Positioning (AP) were recognized mostly as FF as both trajectories had contact with the boundary of the image.

In terms of HMM, one model was trained for each type of exploration procedure. Therefore, in this experiment, five models were trained for classification. During classification, the testing sample was fed into all five models and the probability that this sample belongs to each model was calculated. The label of model with the highest probability was determined as the predicted label. In this study, every model had five hidden states and k-means was applied to categorize the observations into discrete values as the features contain continuous values. The value of k was determined empirically as 10 in this experiment.An accuracy of 28.70% was obtained with the confusion matrix shown in Fig. 7b. Except Frame Following (FF), HMM has difficulties distinguishing the rest procedures.

D. Training SNNs with different dynamics of neurons

The spiking neurons used to build the network were modeled based on the approach proposed by Izhikevich [40] which depends on two variables, the neuron's membrane potential and the membrane recovery variable. There were four parameters a, b, c and d defined in the model that determined the property of the neuron, thus affecting neuron firing activity. Parameter a describes how fast the neuron recovers, where smaller values leads to slower recovery. Parameter b indicates the sensitivity of the recovery variable to fluctuations



(a) (b)

Fig. 7. Confusion matrix of recognition accuracy for (a) DTW with 61.30% accuracy; (b) HMM with 28.70% accuracy.

TABLE IV
RECOGNITION ACCURACY OF DIFFERENT NEURONS.

Neuron Type	a	b	c	d	Accuracy
(a) Tonic Spiking	0.02	0.2	-65	6	94.39%
(b) Phasic Spiking	0.02	0.25	-65	6	72.76%
(c) Tonic Bursting	0.02	0.2	-50	2	82.99%
(d) Phasic Bursting	0.02	0.25	-55	0.05	75.76%
(e) Mixed Mode	0.02	0.2	-55	4	85.55%
(f) Frequency Adaption	0.01	0.2	-65	8	93.72%
(g) Subthreshold Oscillations	0.05	0.26	-60	0	69.21%
(h) Resonator	0.1	0.26	-60	-1	56.03%
(i) Rebound Spike	0.03	0.25	-60	4	61.15%
(j) Rebound Burst	0.03	0.25	-52	0	73.21%
(k) Threshold Variability	0.03	0.25	-60	4	46.75%
(l) Bistability	1	1.5	-60	0	71.65%
(m) DAP	1	0.2	-60	-21	76.26%
(n) Accommodation	0.02	1	-55	4	18.51%

of membrane's potential. c is the reset value of a neuron's membrane potential and d is the reset value of the recovery variable after it is fired. Fourteen primary types of neurons [41] that were applicable in this study were examined. Their parameters are summarized in Table IV with the classification accuracy achieved by the respective SNNs trained with these different types of neurons. Their responses to different input current are also illustrated in Appendix Fig. 10.

It was observed that the type of neuron that fired a train of spikes and adapted its spiking frequency over time exhibited the highest the accuracy of 94.39% and 93.72%. These were very similar to the higher accuracy of 94.55% achieved by using the regular spiking neurons $(a = 0.02, b = 0.2, c = -65, d = 8)$.

V. Discussions

It is observed from the experimental results that with the encoding of SNNs, a relatively good classification accuracy ($> 90\%$) could be achieved by training with only small amount of data. This differed from state-of-art neural networks for spatio-temporal pattern classification, such as 3D-CNN and RNN [9], [42] in that they require large amounts of data. The ability to perform this well with such few observations is a significant advantage considering that data collection is a daunting task with people with disabilities due to the cognitive and physical effort required.

When compared with DTW, SNN showed its advantages in better capturing the underlying structure of the data through feature encoding. Features in different spaces can all be encoded into one single manifold, as the indices of fired neurons. This is crucial when dealing with features that have dramatic differences, such as continuous features versus discrete features. For the DTW compared in this study, procedure RP is mostly recognized as CF since they share common context information, but the shape of the trajectory is not emphasized, because the Euclidean distance used for feature angle is not comparable for the LCS distance used for feature context index.

(a) (b) (c) (d)

Fig. 8. Examples of complete and incomplete exploration procedures.

It is also found that better classification performance was achieved with neurons firing tonic spikes or tonic spikes with frequency adaption. Compared to phasic spikes or burstings (Fig. 10(b)(c)(d)(e), the frequency of tonic spikes may encode the elapsed time between the onsite of input data, thus captured the unique time-series patterns for different classes.

More importantly, our classification framework was capable of early prediction. It was determined from the collected exploration data that users do not always perform a complete procedure. Instead, partial procedures are frequently performed for different reasons. For example, complete CFs are observed when users trace the whole contour of an object to identify its type and shape. However, it is more often applied partially by users to identify their positions. If the user is trying to understand the object's position relative to the left side of the image, it is common practice for users to partially trace the contour of an object in order to reach the uppermost left side of it and then leave from there to detect the left edge of the image. Fig. 8 shows several examples of CF performed by the users, where (a) shows a complete CF and (b), (c) and (d) shows different forms of incompletion.

A large proportion of exploration procedures found in this study were incomplete. Training and collecting all forms of incomplete samples of procedures is time-consuming and not able to be realistically performed. The classification results demonstrated that the proposed SNN-based framework could recognize partial procedures when it was only trained with complete ones. Conversely, this was particularly challenging for DTW and HMM.

Incomplete trajectories could be successfully recognized, because only the first several time frames of a procedure are sufficient for correct prediction. This early prediction property is a desired attribute to develop intelligent interfaces that can proactively offer assistance to the user before determinative errors occur. For example, the sense of distance and direction is degraded when vision is not available. If the users apply procedure Relative Positioning (RP) to measure the distance between two objects, it is crucial for the users to follow the shortest path on the image. An intelligent interface can therefore indicate the position of the shortest path to the users, once it detects partially the performed procedure.

Nonetheless, the findings of this study have to be seen in light of some limitations. The first limitation concerns the data collected from blindfolded participants rather than blind users. Exploration of digital images using a haptic-based multimodal interface is new to both blind and blindfolded community. Both of them needs to go through extensive training practices to get familiar with this new form of interaction. Although no significant differences were found in most tasks in our previous studies when comparing the performance of blind and blindfolded participants [2], it is worthwhile to validate this proposed framework using data from blind participants in the future. Another limitation of this study origins from the limitation of the haptic interface. Different from exploring paper-based images using both of their hands, blind users use a stylus-style haptic controller to explore digital images. Although they only have a single contact point with the image thus made their explorations less efficient, the blind community can have real-time access to images as fast as their sighted peers, while printing paper-based tactile images can take hours. With the proposed framework in this study, more efficient interfaces can be developed to compensate the limitation of single-point interaction and thus improve the performance of haptic-based image exploration.

VI. Conclusions

Individuals who are blind have developed routine exploratory behaviors to facilitate their understanding of visual features, while they are interacting with images. Automatically recognizing these behaviors grants the potential to develop intelligent systems to further assist them in image exploration and enhance their understanding of the images. We proposed a computational framework in this paper that classifies different exploratory behaviors of blind users. The exploratory behaviors are summarized as five different exploration procedures. These procedures consist of various spatio-temporal patterns that are uniquely characterized by rotational, translational and scale invariant features. Numerical features representing the angle of movements and context related to the image features were further encoded through the training of multiple SNNs. The logic feature, referred as "reference switch", was used later for classification without the encoding of a SNN. The SNN-encoded features were then represented by model strings, that are the characteristic responses from the trained SNNs for the input sample. A distance-based classification scheme was applied in this work. We modified the DTW algorithm with a distance function using the length of LCS to compute the differences between model strings. To make the final decision of the predicted label for a sample, DST was integrated in the framework that combines the knowledge obtained from multiple features. Experimental results show that the proposed framework achieves an accuracy of 94.55% for exploration procedure recognition. This framework leads to encouraging future studies that involve the development of intelligent decision-support systems that automatically assist users in understanding image information at the accuracy equaling that of having human assistance.

ACKNOWLEDGMENT

This project is supported by the Partnerships for Innovation program through the National Science Foundation. The authors would also like to thank the State of Indiana through support of the Center for Paralysis Research at Purdue University.

REFERENCES

[1] R. Iglesias, S. Casado, T. Gutierrez, J. Barbero, C. Avizzano, S. Marcheschi, and M. Bergamasco, "Computer graphics access for blind people through a haptic and audio virtual environment," in *Proceedings. Second International Conference on Creating, Connecting and Collaborating through Computing*, oct 2004, pp. 13–18.

[2] T. Zhang, B. S. Duerstock, and J. P. Wachs, "Multimodal Perception of Histological Images for Persons Who Are Blind or Visually Impaired," *ACM Transactions on Accessible Computing*, vol. 9, no. 3, pp. 1–27, jan 2017.

[3] J. Lee, Y. Kim, and G. J. Kim, "Effects of Visual Feedback on Out-of-Body Illusory Tactile Sensation When Interacting With Augmented Virtual Objects," *IEEE Transactions on Human-Machine Systems*, pp. 1–12, 2016.

[4] S. J. Lederman and R. L. Klatzky, "Extracting object properties through haptic exploration," *Acta Psychologica*, vol. 84, no. 1, pp. 29–40, oct 1993.

[5] A. Vinter, V. Fernandes, O. Orlandi, and P. Morgan, "Exploratory procedures of tactile images in visually impaired and blindfolded sighted children: How they relate to their consequent performance in drawing," *Research in Developmental Disabilities*, vol. 33, no. 6, pp. 1819–1831, nov 2012.

[6] S. J. Lederman and R. L. Klatzky, "Hand movements: A window into haptic object recognition," *Cognitive Psychology*, vol. 19, no. 3, pp. 342–368, jul 1987.

[7] I. Puspitawati, A. Jebrane, and A. Vinter, "Local and Global Processing in Blind and Sighted Children in a Naming and Drawing Task," *Child Development*, vol. 85, no. 3, pp. 1077–1090, may 2014.

[8] Z. Cattaneo, T. Vecchi, M. Fantino, A. M. Herbert, and L. B. Merabet, "The effect of vertical and horizontal symmetry on memory for tactile patterns in late blind individuals," *Attention, Perception, & Psychophysics*, vol. 75, no. 2, pp. 375–382, feb 2013.

[9] P. Zhang, X. Wang, W. Zhang, and J. Chen, "Learning SpatialSpectral-Temporal EEG Features With Recurrent 3D Convolutional Neural Networks for Cross-Task Mental Workload Assessment," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 1, pp. 31–42, jan 2019.

[10] X. Tao, "Data clustering via spiking neural networks through spike timing dependent plasticity," *Proceedings of the International Conference on Artificial Intelligence, IC-AI 04*, 2004.

[11] B. Rekabdar, M. M. Nicolescu, M. M. Nicolescu, and S. Louis, "Using patterns of firing neurons in spiking neural networks for learning and early recognition of spatio-temporal patterns," *Neural Computing and Applications*, vol. 28, no. 5, pp. 881–897, may 2017.

[12] T. Zhang, T. Zhou, B. S. Duerstock, and J. P. Wachs, "Image Exploration Procedure Classification with Spike-timing Neural Network for the Blind," in *Proceedings of the 2018 International Conference on Pattern Recognition*, Beijing, China, 2018, pp. 1–6.

[13] S. GHOSH-DASTIDAR and H. ADELI, "SPIKING NEURAL NETWORKS," *International Journal of Neural Systems*, vol. 19, no. 04, pp. 295–308, aug 2009.

[14] B. Rekabdar, M. Nicolescu, M. Nicolescu, and R. Kelley, "Scale and translation invariant learning of spatio-temporal patterns using longest common subsequences and spiking neural networks," in *Proceedings of the International Joint Conference on Neural Networks*, vol. 2015-Septe. IEEE, July 2015, pp. 1–7.

[15] H. Becker, G. Roberts, J. Morrison, and J. Silver, "Recruiting People With Disabilities as Research Participants: Challenges and Strategies to Address Them," *Mental Retardation*, vol. 42, no. 6, pp. 471–475, Dec 2004.

[16] E. E. Pissaloux, R. Velazquez, and F. Maingreaud, "A New Framework for Cognitive Mobility of Visually Impaired Users in Using Tactile Device," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 6, pp. 1040–1051, dec 2017.

[17] F. Leo, E. Cocchi, and L. Brayda, "The Effect of Programmable Tactile Displays on Spatial Learning Skills in Children and Adolescents of Different Visual Disability," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 7, pp. 861–872, jul 2017.

[18] R. Rastogi, T. V. D. Pawluk, and J. Ketchum, "Intuitive Tactile Zooming for Graphics Accessed by Individuals Who are Blind and Visually Impaired," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 21, no. 4, pp. 655–663, jul 2013.

[19] X. Yang, S. Yuan, and Y. Tian, "Assistive Clothing Pattern Recognition for Visually Impaired People," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 2, pp. 234–243, apr 2014.

[20] F. Buonamici, M. Carfagni, R. Furferi, L. Governi, Y. Volpe, F. Buonamici, M. Carfagni, R. Furferi, L. Governi, and Y. Volpe, "Are We Ready to Build a System for Assisting Blind People in Tactile Exploration of Bas-Reliefs?" *Sensors*, vol. 16, no. 9, p. 1361, aug 2016.

[21] A. Stangl, J. Kim, and T. Yeh, "3D printed tactile picture books for children with visual impairments," in *Proceedings of the 2014 conference on Interaction design and children - IDC '14*. New York, New York, USA: ACM Press, 2014, pp. 321–324.

[22] G. J. Williams, T. Zhang, A. Lo, A. Gonzales, D. P. Baluch, and B. S. Duerstock, "3D Printing Tactile Graphics for the Blind: Application to Histology," in *Annual Rehabilitation Engineering Society of North America Conference*, 2014.

[23] C. Colwell, H. Petrie, D. Kornbrot, A. Hardwick, and S. Furner, "Haptic virtual reality for blind computer users," in *Proceedings of the third international ACM conference on Assistive technologies - Assets '98*. New York, New York, USA: ACM Press, 1998, pp. 92–99.

[24] E. F. Wies, M. S. O'Modhrain, C. J. Hasser, J. A. Gardner, and V. L. Bulatov, "Web-based touch display for accessible science education." Springer, Berlin, Heidelberg, 2001, pp. 52–60.

[25] W. Yu and S. Brewster, "Multimodal virtual reality versus printed medium in visualization for blind people," in *Proceedings of the fifth international ACM conference on Assistive technologies - Assets '02*. New York, New York, USA: ACM Press, 2002, p. 57.

[26] E. Tuominen, M. Kangassalo, P. Hietala, R. Raisamo, and K. Peltola, "Proactive Agents to Assist Multimodal Explorative Learning of Astronomical Phenomena," *Advances in Human-Computer Interaction*, vol. 2008, pp. 1–13, jan 2008.

[27] N. Kaklanis, K. Votis, and D. Tzovaras, "Open Touch/Sound Maps: A system to convey street data through haptic and auditory feedback," *Computers & Geosciences*, vol. 57, pp. 59–67, aug 2013.

[28] Y. Hatwell, A. Streri, and E. Gentaz, *Touching for knowing : cognitive psychology of haptic manual perception*. John Benjamins Pub, 2003.

[29] S. J. Lederman and R. L. Klatzky, "Haptic classification of common objects: Knowledge-driven exploration," *Cognitive Psychology*, vol. 22, no. 4, pp. 421–459, oct 1990.

[30] P. W. Davidson, "Haptic judgments of curvature by blind and sighted humans." *Journal of Experimental Psychology*, vol. 93, no. 1, pp. 43–55, 1972.

[31] S. Ungar, M. Blades, and C. Spencer, "Mental rotation of a tactile layout by young visually impaired children," Tech. Rep., 1995.

[32] F. Gaunet, J.-L. Martinez, and C. Thinus-Blanc, "Early-Blind Subjects' Spatial Representation of Manipulatory Space: Exploratory Strategies and Reaction to Change," *Perception*, vol. 26, no. 3, pp. 345–366, mar 1997.

[33] E. M. Izhikevich, "Polychronization: Computation with Spikes," *Neural Computation*, vol. 18, no. 2, pp. 245–282, feb 2006.

[34] ——, "Which Model to Use for Cortical Spiking Neurons?" *IEEE TRANSACTIONS ON NEURAL NETWORKS*, vol. 15, no. 5, p. 1063, 2004. [Online]. Available: www.izhikevich.com.

[35] M. Arif, T. Brouard, and N. Vincent, "A fusion methodology based on Dempster-Shafer evidence theory for two biometric applications," in *Proceedings - International Conference on Pattern Recognition*, vol. 4, 2006, pp. 590–593.

[36] X. Li, A. Dick, C. Shen, Z. Zhang, A. van den Hengel, and H. Wang, "Visual tracking with spatio-temporal DempsterShafer information fusion," *IEEE Transactions on Image Processing*, vol. 22, no. 8, pp. 3028–3040, 2013.

[37] T. Zhou and J. P. Wachs, "Early Turn-taking Prediction with Spiking Neural Networks for Human Robot Collaboration," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, p. (in press).

[38] "Dynamic Time Warping," in *Information Retrieval for Music and Motion*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 69–84.

[39] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[40] E. M. Izhikevich, "Dynamical Systems in Neuroscience:," pp. 227–256, 2004.

[41] ——, "Which Model to Use for Cortical Spiking Neurons?" *IEEE TRANSACTIONS ON NEURAL NETWORKS*, vol. 15, no. 5, p. 1063, 2004.

[42] J. Sanchez, D. Erdogmus, M. Nicolelis, J. Wessberg, and J. Principe, "Interpreting Spatial and Temporal Neural Activity Through a Recurrent Neural Network BrainMachine Interface," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 13, no. 2, pp. 213–219, jun 2005.